DEPARTAMENTO DE COMPUTACION

Facultad de Ciencias Exactas y Naturales - UBA

TP 2: Análisis de Redes Sociales

October 16, 2022 Métodos Numéricos

Grupo 18

Integrante	LU	Correo electrónico
Vekselman, Natán	338/21	natanvek11@gmail.com
Arienti, Federico	316/21	fa.arienti@gmail.com

Facultad de Ciencias Exactas y Naturales

Universidad de Buenos Aires

Ciudad Universitaria - (Pabellón I/Planta Baja)

Intendente Güiraldes 2610 - C1428EGA

Ciudad Autónoma de Buenos Aires - Rep. Argentina

Tel/Fax: (++54 +11) 4576-3300

http://www.exactas.uba.ar

RESUMEN

La descomposición de matrices en autovectores y autovalores aparece en una variedad de aplicaciones donde importa caracterizar el comportamiento de un sistema: en el reconocimiento de imágenes, en el análisis de estabilidad de cuerpos rotantes, en el análisis de riesgo de mercado y en el análisis de redes —por nombrar algunas—. Desde un punto de vista geométrico, se puede considerar a los autovectores como los 'ejes' de una transformación lineal, en tanto representan una dirección invariante a la transformación, y a los autovalores como los factores por los que esas direcciones se comprimen, estiran o invierten.

En este trabajo propondremos una implementación en C++ de un método para el cálculo de autovalores reales, no nulos y en módulo dominantes, y sus respectivos autovectores, en matrices cuadradas. Para algunas matrices particulares, como pueden ser las matrices simétricas definidas positivas, este método nos permitirá obtener todos sus autovalores y autovectores asociados. El mismo se conoce como *el método de la potencia con deflación*.

A su vez, presentaremos dos aplicaciones concretas de los autovalores y autovectores en el análisis de redes: la medición de centralidad de autovector y corte mínimo en la red del *'Club de Karate'* [3], y la estimación de una *ego-network* [2] de Facebook, por medio de la construcción de una matriz de similaridad.

Palabras clave: *método de la potencia, deflación de Hotelling, centralidad de autovector, conectividad algebráica, análisis de componentes principales.*

CONTENIDOS

1. Método de la potencia con deflación

1.1. **Introducción teórica.** El método de la potencia con deflación permite aproximar un subconjunto de los autovalores y autovectores asociados a una matriz. Si la misma satisface que todos sus autovalores son no nulos y diferentes en módulo, entonces permite aproximar el conjunto entero.

Método de la potencia: el método de la potencia, *Power method* o *Power iteration*, es una técnica iterativa para aproximar el autovector asociado al autovalor en módulo máximo de una matriz cuadrada que satisfaga esta característica —es decir, tenga un autovalor dominante no nulo—, a partir de la aplicación de sucesivos productos matriciales, descriptos por la siguiente relación de recurrencia:

$$b_0 \text{ es un vector aleatorio} : ||b_0|| = 1$$

$$b_{k+1} = \frac{\mathbf{A} b_k}{||\mathbf{A} b_k||} \tag{1}$$

donde $||\cdot||$ es una norma vectorial.

Se puede demostrar [1] que, bajo las condiciones descriptas, si b_0 no es ortogonal al autovector asociado al autovalor dominante en módulo de $\mathbf{A}$, b_k convergerá a éste. Lo que es más, se podrá aproximar el autovalor dominante por medio del coeficiente de Rayleigh:

$$\lambda_{max} = \frac{b_k^t \ \mathbf{A} \ b_k}{b_k^t \ b_k} \tag{2}$$

Método de la deflación: el método de la deflación, por su parte, corresponde a la transformación de la matriz inicial $\mathbf{A}$ por una matriz $\mathbf{B}$ con autovalores equivalentes, salvo por el autovalor dominante que será anulado. Existen distintos métodos de deflación, entre ellos la deflación de Hotelling y la deflación de Wielandt [1].

En este trabajo utilizarmos la deflación de Hotelling por su sencillez, a cuestas de un mayor error numérico [1]. El mismo consiste en aplicar el método de la potencia para sucesivas matrices que satisfagan la siguiente relación de recurrencia:

$$\mathbf{B}_0 \ = \mathbf{A}$$

$$\mathbf{B}_{k+1} = \mathbf{B}_k - \lambda \ v \ v^t \tag{3}$$

donde λ corresponde al autovalor en módulo máximo estimado por el método de la potencia y v a su autovector asociado.

Se puede demostrar [citar] que $\mathbf{B}_{k+1}$ contiene los mismos autovalores que $\mathbf{B}_k$, salvo el máximo que quedará anulado.

1.2. **Implementación.** Procedemos a detallar una posible implementación para ambos métodos, restringiéndonos al caso de autovalores reales. Definimos:

$$\mathit{deflacion} : \mathit{matriz}_{n\times n}\ \mathbf{A} \times \mathit{nat}\ \mathrm{k} \times \mathit{nat}\ \mathrm{q} \times \mathit{real}\ \mathrm{t} \longrightarrow \mathit{vector}_k \times \mathit{matriz}_{n\times k}$$

$$\mathit{potencia} : \mathit{matriz}_{n\times n}\ \mathbf{A} \times \mathit{nat}\ \mathrm{q} \times \mathit{real}\ \mathrm{t} \longrightarrow \mathit{real} \times \mathit{vector}_n$$

donde n es un natural, $\mathbf{A}$ tiene al menos k autovalores reales dominantes en módulo, $0 < k \leq n$, q es un número par[1] que representa el máximo de iteraciones a realizar y $0 \leq t$ representa la tolerancia mínima a partir de la que se considera la convergencia de una solución.

```
1  proc deflacion(in A: matriz<n, n>, in k: Nat, in q: Nat, in t: Real) {
2
3      eigvals := vector<q>
4      eigvecs := matriz<n, k>
5
6      i := 0
7      while i < k {
8
9          l, V := potencia(A, q, t)
10         eigvals[i] := l
11         eigvecs.columna[i] := V
12
13         A := A - l * (V * V.t)
14         i := i + 1
15     }
16
17     return eigvals, eigvecs
18 }
```

Algoritmo 1. Pseudocódigo para el método de la deflación.

El algoritmo (1.) retornará un vector con los primeros k autovalores en módulo máximos de $\mathbf{A}$, ordenados descendientemente, y una matriz cuyas columnas corresponden, respectivamente, a los autovectores normalizados asociados a estos autovalores.

Es intersante notar que la k-ésima matriz sobre la que se aplicará el método de la potencia —$\mathbf{B}_k$— tendrá, por definición, un autovalor cero con multiplicidad algebráica mayor o igual

[1] Esta restricción no es necesaria, pero permite realizar una optimización que se explica en la próxima página.

a k. En el caso en que la matriz inicial sea simétrica, $\mathbf{B}_k$ será simétrica[2] y, en consecuencia, diagonalizable. Se puede demostrar que la única matriz diagonalizable con multiplicidad algebráica $\mu_a(0) = n$ es la matriz nula, por lo que el método de la deflación de Hotelling tenderá hacia ella. Esto nos permite inferir que el error númerico será proporcional a k[3]. Es decir, los autovalores más chicos de la matriz serán más susceptibles a errores.

Por su parte, el algoritmo (2.) retornará el autovalor de $\mathbf{A}$ máximo en módulo y su autovector asociado:

```
proc potencia(in A: matriz<n, n>, in q: Nat, in t: Real) {

    q := q / 2
    B := A * A

    v := aleatorio(n)      // un vector aleatorio no nulo
    v := v / norma(v, 2)   // ||v||_2 = 1

    i := 0
    while i < q {

        y := B * v
        y := y / norma(y, 2)
        if (norma(v - y, 2) < t) {  // tolerancia para la convergencia
            break
        }
        v := y
        i := i + 1
    }

    l := (v.t * A * v) / (v.t * v)  // coeficiente de rayleigh

    return l, v
}
```

Algoritmo 2. Pseudocódigo para el método de la potencia.

Una primera observación importante es que el algoritmo (2.) no es capaz de distinguir si la selección inicial del vector v es ortogonal al autovector asociado al autovalor en módulo dominante, por lo que una mala selección puede resultar en que el algoritmo falle. Para reducir la probabilidad de ocurrencia, se propone —heurísticamente— que se elija cada coordenada del vector de manera pseudo-aleatoria sobre un rango amplio. Por ejemplo, la máxima representación de enteros con signo en 32 bits.

[2]Por suma de simétricas. Notar que vv^t siempre resulta en una matriz de este tipo.

[3]En tanto existirá una correlación. Sin embargo, es esperable que otros factores entren en juego: la variabilidad de los autovalores, el número de condición de la matriz, la selección del vector aleatorio inicial, o la cantidad de iteraciones q a realizar, por ejemplo.

Además, proponemos un modelo diferente al que define la relación de recurrencia de la ecuación (1.). Su convergencia, en realidad, depende del signo del autovalor en módulo dominante. Si este es negativo, la secuencia será acotada. Se puede demostrar[4] que la subsecuencia definida por $\{b_k \ : \ k \text{ es par}\}$ siempre convergerá.

Lo que es más, esta subsecuencia permite que el método funcione para matrices con distintos autovalores en módulo dominantes, en tanto estos no sean nulos[5]. El algoritmo (2.) aprovecha estas observaciones.

Desde un punto de vista temporal, también permite reducir a la mitad el costo de ejecución (se itera $q/2$ veces).

1.3. **Evaluación cuantitativa.** Procedemos a evaluar nuestra implementación del método de la potencia con deflación en C++ acorde a los algoritmos propuestos.

Error relativo: medimos el error $|\mathbf{AV} - \mathbf{V\Lambda}|_1$ en función de la cantidad de iteraciones q para 300 instancias de matrices $\mathbf{A} \in \mathbb{R}^{20\times 20}$ generadas aleatoriamente, donde $\mathbf{V}$ y $\mathbf{\Lambda}$ representan —respectivamente— las matrices aproximadas de autovectores y autovalores de $\mathbf{A}$, tal que $\mathbf{AV}_i \approx \mathbf{\Lambda}_{ii}\mathbf{V}_i \quad \forall i : \ 1 \ ... \ n$. En total, obtuvimos tres millones de mediciones[6].

Metodología: se calculó $\mathbf{\Lambda}, \mathbf{V} = \textit{deflacion}(\mathbf{A}, \ 20, \ q, \ 0)$ y se midió el error relativo para cada una de las matrices sobre cada valor de q en en el intervalo $(0, 1e5)$, de a saltos de 10.

Cada caso se generó a través de uno de los siguientes tres procedimientos[7]:

1) *Matrices Diagonales*: Se generaron cien matrices diagonales $\mathbf{D}$ con veinte autovalores en el rango $[-1e5, 0) \cup (0, \ 1e5]$. Los mismos se generaron con el rng *PCG64* de numpy para evitar distribuciones particulares que pudieran influir en la variabilidad de los autovalores.

2) *Matrices Diagonalizables*: Se generaron cien matrices diagonalizables $\mathbf{A} = \mathbf{QDQ}^t$ donde cada matriz $\mathbf{D}$ se generó a partir de la metodología (1.) y $\mathbf{Q} = \mathbf{I} - 2uu^t$ se generó a partir de un vector aleatorio u —con el algoritmo random.rand() de numpy— tal que $||u||_2 = 1$.

3) *Matrices Simétricas Definidas Positivas*: Se generaron cien matrices simétricas definidas positivas de enteros. Para cada caso se generó una matriz aleatoria $\mathbf{B}$ con el algoritmo random.randint() de numpy y se procedió a definir la matriz $\mathbf{A} = \mathbf{BB}^t$.

[4] [demostrar]
[5] [demostrar]
[6] El script asociado se puede encontrar en *./experimentos/error_potencia.py*
[7] Se utilizó un valor semilla para facilitar la reproductibilidad.

Observaciones: consideramos que la variabilidad de los autovalores y el número de condición de las matrices son variables que afectan de manera significativa al error relativo. El proceso mencionado para la generación de matrices aleatorias fue pensado sobre generadores de números aleatorios para tratar de mantener a ambas variables controladas, en tanto no respondan a ninguna distribución particular que pueda exhibir sus carecterísticas en los resultados del experimento.

Resultados. [experimento]

2. Análisis: Club de Karate

2.1. **Contexto.** La red del *Club de Karate* es parte de una investigación antropológica [citar] que estudió las relaciones 'políticas' entre los miembros de un club universitario. La misma se realizó durante el desarrollo de un conflicto que terminó por dividir al grupo. La red busca modelar el flujo de información entre sus integrantes por medio de la tripla ($\mathbf{V}$, $\mathbf{E}$, $\mathbf{C}$), donde $\mathbf{V}$ denota el conjunto de individuos, $\mathbf{E}$ refiere a un grafo no direccionado cuyos ejes representan los vínculos entre los miembros, y $\mathbf{C}$ define la fuerza de estas relaciones —lo que se podría pensar como ponderaciones sobre los ejes de $\mathbf{E}$—.

La investigación tuvo como eje central demostrar la capacidad del modelo para predecir la división del grupo por medio del algoritmo de *labeling* de 'flujo máximo - corte mínimo' de Ford y Fulkelson [citar] .

En este análisis utilizaremos sólo la representación matricial de $\mathbf{E}$ para evaluar la importancia de los distintos miembros en la red, y la matriz laplaciana asociada para evaluar el uso de autovectores como predictores de la división del grupo.

2.2. **Centralidad de Autovector.** La centralidad de autovector es una medida que se utiliza en el análisis de redes para medir la 'importancia' de los nodos que componen una red, relativa a la importancia de sus conexiones. Dada una matriz de conectividad $\mathbf{W}$, se define:

$$\lambda x = \mathbf{W}x \tag{4}$$

donde λ es el autovalor en módulo máximo de $\mathbf{W}$.

Intuitivamente, se puede pensar que la importancia de cada nodo es proporcional a la suma de las importancias de sus vecinos [citar] . Se puede demostrar [citar] que, dado las características de esta matriz, el autovector asociado siempre será positivo en coordenadas.

En tanto la red de *Club de Karate*, podemos ver que la aplicación del metodo de la potencia con deflación sobre la matriz de conectividad asociada al grafo $\mathbf{E}$ resulta en el siguiente autovector asociado al autovalor en módulo máximo:

[presentar vector.]

Vemos que el nodo '1' y el nodo '34' son los más 'centrales'. Esto no es casualidad, la red del *Club de Karate* está armada para tener a las dos figuras centrales del conflicto en cada extremo —el instructor de karate y el presidente del club—, para satisfacer la especificación del algoritmo de 'labeling' que utiliza.

2.3. **Autovectores de la matriz laplaciana.** La matriz laplaciana $\mathbf{L} = \mathbf{D} - \mathbf{W}$ —donde $\mathbf{D}$ es una matriz diagonal con elementos $d_{ii} = \sum_j w_{ij}$ y $\mathbf{W}$ es una matriz de conectividad—

sirve para medir distintas propiedades en una red. En particular, el mínimo autovalor en módulo no nulo —llamado de conectividad algebráica, o valor de Fiedler— permite establecer un criterio sobre el que particionar la red en dos. El autovector asociado a este autovalor designará la pertenencia de un nodo a una u otra partición acorde a su signo.

Procedemos a analizar qué autovector de la matriz laplaciana asociada a $\mathbf{E}$ permite predecir mejor la división que ocurrió en el *Club de Karate*. Para ello, medimos el valor absoluto de la correlación entre cada autovector y un vector que indica la división que ocurrió en el grupo.

[presentar solución.]

3. Análisis: Red 'Ego'

3.1. **Contexto.** Una *ego-network* [citar] es una red compuesta por las amistades que existen entre los amigos de un individuo, el 'ego'. Estas redes son centrales para aplicaciones como Facebook, Google+ o Twitter. En particular, dada una red 'ego', resulta de interés poder identificar los circulos sociales —conjuntos, disjuntos y anidados— a los que pertenece un usuario. En [citar] se propone un método de aprendizaje no supervisado para lograr inferirlos, que se nutre de la siguiente información: un grafo **E** —la red—, donde se espera que exista una correlación fuerte entre un círculo y la densidad de conexiones entre los nodos que lo componen; y un conjunto de atributos **C**, para cada nodo, donde se espera una correlación entre un círculo y la similaridad de los atributos de los nodos que lo componen—.

En este análisis estimaremos **E**, la red 'ego' original, por medio de la construcción de una matriz de similaridad que utilice los atributos definidos en **C**. También, buscaremos reducir su dimensionalidad por medio del análisis de componentes principales.

3.2. **Matriz de similaridad.**

3.3. **Comparación con la red original.**

3.4. **Optimización.**

3.5. **PCA.**

4. Conclusiones

5. Apéndice

Referencias

[1] Richard L Burden and J Douglas Faires. *Numerical Analysis*. 2000.

[2] Jure Leskovec and Julian Mcauley. Learning to discover social circles in ego networks. *Advances in neural information processing systems, 25*, 2012.

[3] Wayne W Zachary. An information flow model for conflict and fission in small groups. *Journal of anthropological research, 33(4):452–473*, 1977.